Serial # 09/724 182
Applicant: R. E. M. SWANN, et al
Title: Image Processing for Clustering Related Text Objects

The Patent Office

INVESTOR IN PEOPLE

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed

Dated 11 DEC 2000

Patents Form 1/77

Patents Act 1977
(Rule 16)

THE PATENT OFFICE
30 JUN 2000
RECEIVED BY HAND

The Patent Office

1/77

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form.)

The Patent Office

Cardiff Road
Newport
Gwent NP9 1RH

Fee: £0

1.	Your reference	AJR/41610
2.	Patent application number *(The Patent Office will fill in this part)*	**0016219.8**
3.	Full name, address and postcode of the or of each applicant *(underline all surnames)*	THE POST OFFICE 148 Old Street London EC1V 9HQ ENGLAND
	Patents ADP number *(if you know it)*	501260440006
	If the applicant is a corporate body, give the country/state of incorporation	British
4.	Title of the invention	IMAGE PROCESSING FOR CLUSTERING RELATED TEXT OBJECTS
5.	Full name, address and postcode in the United Kingdom to which all correspondence relating to this form and translation should be sent	Reddie & Grose 16 Theobalds Road LONDON WC1X 8PL
	Patents ADP number *(if you know it)*	91001

30 JUN 2000

03JUL00 E549635-9 D01631
~~P01/7700 0.00-0016219.8~~

		Country	Priority application *(If you know it)*	Date of filing *(day/month/year)*
6.	If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number			

		Number of earlier application	Date of filing *(day/month/year)*
7.	If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application		

8.	Is a statement of inventorship and of right to grant of a patent required in support of this request? *(Answer 'Yes' if:* *a) any applicant named in part 3 is not an inventor, or* *b) there is an inventor who is not named as an applicant, or* *c) any named applicant is a corporate body.* *See note (d))*	YES

9. Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document.

Continuation sheets of this form	
Description	8
Claim(s)	3
Abstract	1
Drawing(s)	5 ~~1~~

10. If you are also filing any of the following, state how many against each item.

Priority documents	
Translations of priority documents	
Statement of inventorship and right to grant of a patent *(Patents Form 7/77)*	
Request for preliminary examination and search *(Patents Form 9/77)*	1
Request for substantive examination *(Patents Form 10/77)*	
Any other documents *(please specify)*	

11. I/We request the grant of a patent on the basis of this application.

Signature

Date
30 June 2000

12. Name and daytime telephone number of person to contact in the United Kingdom

A J ROBSON
020-7242 0901

Warning
After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or such direction has been revoked.

Notes

a) If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.

b) Write your answers in capital letters using black ink or you may type them.

c) If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.

d) If you have answered 'Yes' Patents Form 7/77 will need to be filed.

e) Once you have filled in the form you must remember to sign and date it.

f) For details of the fee and ways to pay please contact the Patent Office.

IMAGE PROCESSING FOR CLUSTERING RELATED TEXT OBJECTS

This invention relates to image processing and is of particular benefit in processing document images for purposes such as text identification. This type of process is particularly useful in address identification for automated reading of addresses in mail processing systems but is generally applicable to the text processing of document images.

In processing a document image a common requirement is to isolate text objects and then cluster them into blocks of text (e.g. paragraphs, addresses, form fields etc). The first step is to extract the text objects resulting in a binary image. Clustering the binary text objects together then produces text blocks. A simple clustering technique would be to merge any neighbouring binary objects where the gap between the objects is less than a given limit. In documents where the blocks of text of interest are well separated from other text blocks, this technique is usually successfully. For example, for a simple white envelope with an address and a single stamp or indicia, the distance between letters in the address is usually significantly less than the distance between the address and the other objects on the envelope. Thus it is easy to set a distance limit in the merging that will cluster together the text objects of the address but not connect the address to the stamp or indicia which is significantly further away. However, on more complex document images, text block identification is not so simple. For example, many mail items have additional text and designs printed on them. Also, some mail items come in transparent packaging which allows the content to be seen from the outside. It is a more difficult task to identify likely candidates for an address from images of these types of documents.

Complex document images often require different processing to the more simple images to identify and read, e.g. address information. Accordingly in the case of processing mail we have proposed in our British Patent Application No. [ ] filed on the same day as the current application, a method and apparatus for identifying the degree of complexity in an image of a mail item, so that subsequent processing is appropriate to the image.

The merging of binary text objects from a document image connects together text objects that have a gap between them that is less than a distance limit. The merging distance limit needs to be appropriate to the text. If the distance limit is too small then some text within a block remains un-merged and isolated. If the distance limit is too large then over-merging occurs where distinct text blocks are incorrectly linked. The appropriate merging distance limit is usually proportional to the font size of the text in the document being processed. In simple documents where the text block of interest is well separated from other objects in the image the setting of the merging distance limit is easy. However, in complex image documents there may be text objects of a variety of fonts, sizes and orientations. With such images the clustering of text objects may not be possible with a single merging distance limit, since a limit which is too small for some text objects in the document may be too large for text in another part of the image. In such situations the clustering of text objects into text blocks usually becomes more complex as techniques should connect together only text of the same font, orientation, etc.

We have appreciated that the process of clustering related text objects can be made more computationally efficient and successful by approaching it as a segmentation problem followed by a simple merging routine, rather than

investing in a complicated or computationally expensive clustering technique. Accordingly, a preferred embodiment of the present invention first segments the binary image of the text objects in such a way as to separate text objects that are close but unrelated. This effectively separates the complex image into separate simple images where a coarse computationally light merge can be applied to successfully cluster the text objects into relevant text blocks. In a preferred embodiment of the invention, the information used to segment the binary text objects is taken from the original greyscale or colour document image. Items of different greyscale or colour value are assumed to be unrelated. Items on different greyscale or colour value are assumed to be unrelated.

There are a number of local text attributes that generally remain constant for a text block in a document image. The useful measures that could be extracted from the binary text objects such as font, orientation and alignment are usually computationally demanding to calculate. The local text attribute used in this preferred embodiment is a local measure of the text's original colour or greyscale value. This is computationally easy to measure but requires the interrogation of the original image. This is different to most document image processing techniques, which discard the original image information once the binary text objects have been extracted. In addition this invention also uses a global document image attribute to separate unrelated text blocks. This is the background colour in the document, which again in standard techniques would have been lost in discarding the original image. These measures are useful as long as the assumptions that text blocks are printed in a consistent colour and on a consistent background hold for the document image being processed.

The invention is defined with more precision in the appended claims to which reference should now be made.

A preferred embodiment of the invention will now be described in detail by way of example with reference to the accompanying drawings in which:

Figure 1 shows schematically apparatus for identifying addresses on envelopes;
Figure 2 shows schematically an embodiment of the invention for isolating text blocks in a document image;
Figure 3 shows schematically how segmentation from global background information is used to assist in the creation of clusters of related text objects;
Figure 4 shows schematically how segmentation from local minimum greyscale information is used to assist in the creation of clusters of related text objects; and
Figure 5 shows schematically an embodiment of the invention being used in the example of identifying addresses on complex envelopes

In the processing of document images the first stage is to acquire the image. This image acquisition may be through the use of a scanner camera or other interface. In the example of a mail processing system the resultant document image is a 256 level greyscale image. Embodiments of the invention may be designed for original document images that are greyscale or colour, but not simple binary images.

In the example of a mail processing system, the stages of identifying addresses on envelopes, as will be used in an embodiment of the invention, are shown in Figure 1. A scanner 2 produces a 256 level greyscale image in which each picture element (pixel) has assigned to it a 256 level greyscale value representing the darkness of the

image at that pixel. A text object extraction unit 4 then produces a binary image of text objects within the image. The text object extraction unit operates according to one of a number of methods of operation which will be familiar to those skilled in the art. This binary data defines the size and location of all objects that may be letters and passes it to a clustering unit 6. This creates blocks of related text objects which are passed to a sorting unit 8 which orders the text blocks according to their properties (e.g., number of lines, number of letters, location on the envelope). This sorts the text blocks in order of decreasing likeliness of being an address and then sends these in turn to an optical character recognition device (OCR) which can read the text blocks.

With simple white envelopes the clustering unit in the mail processing system can be a merger routine which removes spaces smaller than a predetermined size between the text objects. Thus, where the objects are letters it should merge them into paragraphs. In the case of simple white envelopes the determining of the merger distance limit is easy because the text blocks are well separated. In more complex documents it is not always possible for a single merger distance limit to successfully cluster all the text blocks without also over-merging some of the data (i.e. connecting some unrelated text blocks).

Embodiments of the invention cluster together related text objects into paragraphs in cases where a simple merger is not sufficient. The embodiment, shown in Figure 2, uses a text object extraction unit 4 as before to extract the binary text data. However the original image provided by the scanner is also retained and used to provide two different segmentations of the image.

The first segmentation is performed by a local information segmentation unit 12. This receives data from the text

object extraction unit indicating which objects are text in the image and also receives data from the original greyscale/colour image to give a measure of the text greyscale or colour for each item of text. Once the greyscale/colour has been calculated for all the text objects, a segmentation of the image is created from the data. For example, in a mail processing system a measure of a minimum grey level of the letters is used as the local measure of text colour. The segmentation created by the local information segmentation unit 12 segments blocks of text of similar colour together. Thus, if items of text are within the normal merging limit of the clustering system but are of different greyscale or colour values, they will fall into different text blocks and cannot subsequently be merged.

The second segmentation is a global information segmentation unit 14. This performs a segmentation based on the global information in the background of the original image. The unit generates a background image from the original greyscale/colour image and looks for differences in background colour/greyscale to perform segmentations between areas of different background greyscale/colour.

Data from the text object extraction unit 4, the local information segmentation unit 12, and the global information segmentation unit 14 are then provided to a master segmentation unit 16. This applies the segmentation information provided by the local and global segmentation units to the binary image of the text objects. Thus, the text objects of a single complex image are converted to the text objects of a plurality of simpler segmented regions. Each of these regions contains text objects where the background and text greyscale/colour are constant. The data for each region is then supplied in turn to a merging unit 18 which merges

text objects for each region into clusters of related text objects. This results in groups of text blocks. From here they can be provided to the sorting unit 8 of Figure 1 and subsequently to the optical character recognition unit 10. This enables the most likely candidates for the address on the envelope to be identified and read.

Figure 3a)-f) shows diagrammatically how text objects in a document which would normally be clustered together using a simple merger unit are correctly clustered into related text blocks when each segmented background region is processed separately using apparatus from that of Figure 2. Figure 3a) shows the original document image and Figure 3b) the binary image produced by the text object extraction unit 4. Figure 3c) shows the results of clustering which would be produced by the arrangement of Figure 1. Figure 3d), however, shows the data which would be extracted by the global information segmentation unit 14 of Figure 2. In this the background regions are clearly identified. Figure 3e) then shows how the segmentation unit 16 segments the image according to the different background regions and how one of those regions then has its text objects merged by the merging unit 18. Figure 3f) shows the results and clusters of text objects which have consistent backgrounds which can then subsequently be sorted to determine the most likely candidates for address blocks.

In Figure 4, the use of the minimum grey level as a local text colour measurement as performed by the local information segmentation unit 12 is shown. The original document image is shown in Figure 4a). In Figure 4b) the binary image produced by text object extraction unit 4 and the original document image are combined and the minimum grey level for each text object is obtained from the original document image. The local minimum grey level is a measure of text colour or darkness. Thus, it can be

seen that there are several different grey levels in Figure 4b).

The segmentation then produced by local information segmentation unit 12 is shown in Figure 4c) where the image is divided into a number of regions, each of which have the same text colour or greyscale. This information is fed into the segmentation unit 16 for use in providing the overall segmentation which is input to the merging unit 18.

Figure 5 shows an application embodying the invention of Figure 2 in which the sorting unit 8 and optical character recognition unit 10 are shown receiving the output from the merging unit 18. In this particular example, the scanner 2 produces a 256 level greyscale image. Clearly, it could also produce a colour image. Other than this, the apparatus operates in the same manner as that of Figure 2. As a mail processing system, it is able to deal with document images of more complex envelopes than has previously been the case.

CLAIMS:

1. An image processing system for producing clusters of related objects for subsequent analysis comprising, means for supplying a multi-level digital representation of the image, means for identifying predetermined objects in the image and supplying data defining their locations, means for deriving segmentation data from the multi-level digital representation, means for segmenting the data defining the locations of objects with the said segmentation data, means for clustering the predetermined objects into groups for each segmented region and means supplying data relating to the groups for subsequent analysis.

2. An image processing system according to claim 1 comprising second segmentation means receiving the multi-level representation of the image and the data identifying the locations of the objects in the image to derive segmentation data relating to differences in the data of the multi-level representation from the said objects, and wherein the segmenting means receives this segmentation data and further segments the data defining locations of object image in dependence on this segmentation data.

3. An image processing system according to claim 1 or 2 which the objects are text objects.

4. An image processing system according to claim 1, 2 or 3 in which the means for supplying segmentation data from the multi-level representation comprises means for comparing differences in background data to derive segmentation data.

5. An image processing system according to claim 4 in which the background data comprises colour data.

6. An image processing system according to claim 4 in which the background data comprises greyscale data.

7. An image processing system according to any of claims 2-6 in which the second segmentation means derives segmentation data from the colour of the objects.

8. An image processing system according to any of claims 2-6 in which the second segmentation means derives segmentation data from the greyscale level of the objects.

9. A method for processing images to produce clusters of related objects for subsequent analysis comprising the steps of supplying a multi-level digital representation of the image, identifying predetermined objects in the image, supplying data defining the locations of these predetermined objects, deriving segmentation data for the image from the multi-level digital representation, segmenting the data defining the locations of the objects with the segmentation data, clustering the objects into groups for each segmented region, and supplying data relating to the groups for subsequent analysis.

10. A method according to claim 9 further comprising the steps of receiving the multi-level digital representation of the image and the data identifying the location of the objects in the image and deriving additional segmentation data relating to differences in the data of the multi-level representation for the said objects, and wherein the segmenting step further segments the data defining the locations of objects in dependence on this additional segmentation data.

11. A method according to claims 9 or 10 in which the objects are text objects.

12. A method according to claim 9 or 10 in which the step of supplying segmentation data for the multi-level digital representation comprises comparing differences in background data to determine segmentation data.

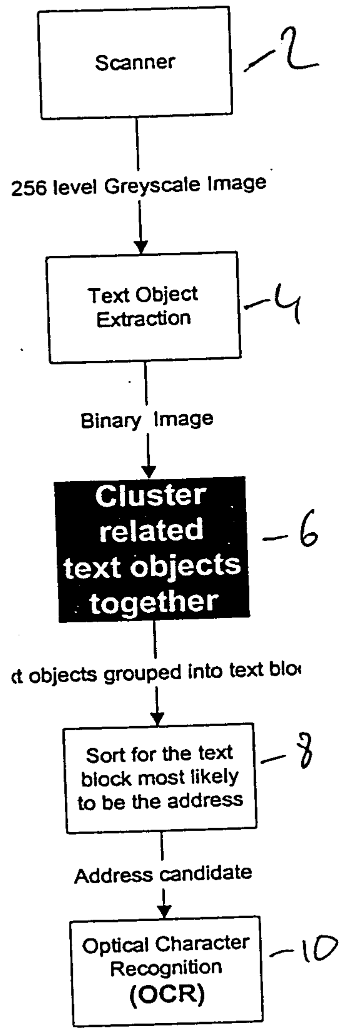
13. A method according to claim 11 in which the background data comprises colour data.

14. A method according to claim 12 in which the background data comprises greyscale data.

15. An image processing system for producing clusters of related objects for subsequent analysis substantially as herein described with reference to figures 2-5 of the drawings.

16. A method for processing images to produce clusters of related objects for subsequent analysis substantially as herein described.

Figure 1.


Scanner
2
256 level Greyscale Image
Text Object Extraction
4
Binary Image
Cluster related text objects together
6
xt objects grouped into text blo
Sort for the text block most likely to be the address
8
Address candidate
Optical Character Recognition (OCR)
10

2|5

Figure 2.

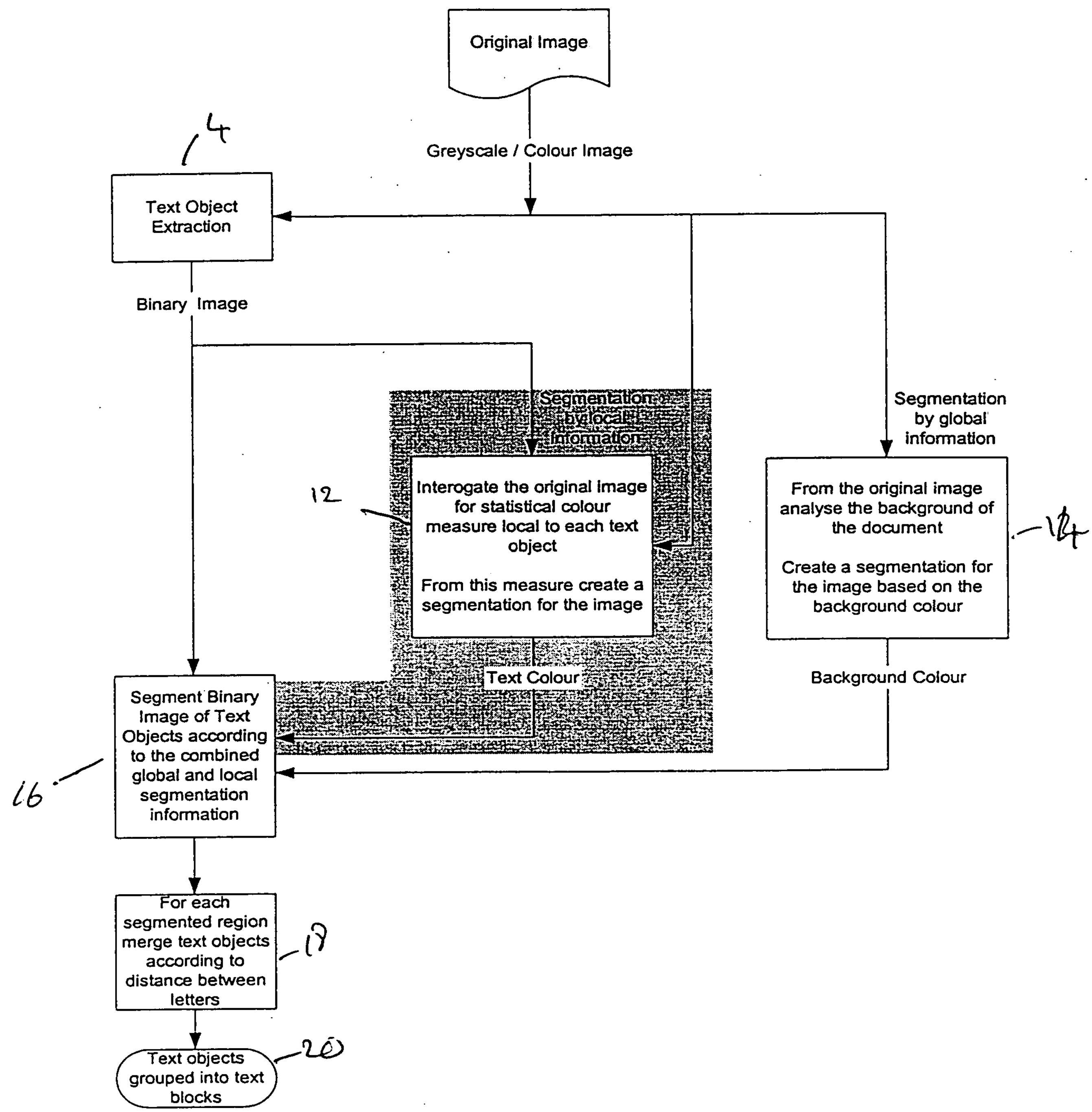
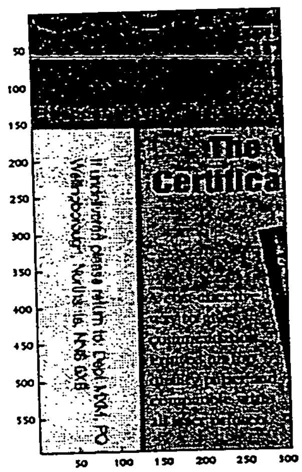

Original Image
Greyscale / Colour Image
Text Object Extraction
4
Binary Image
Segmentation by local information
Interogate the original image for statistical colour measure local to each text object
From this measure create a segmentation for the image
12
Segmentation by global information
From the original image analyse the background of the document
Create a segmentation for the image based on the background colour
14
Text Colour
Background Colour
Segment Binary Image of Text Objects according to the combined global and local segmentation information
16
For each segmented region merge text objects according to distance between letters
18
Text objects grouped into text blocks
20

Figure 3.

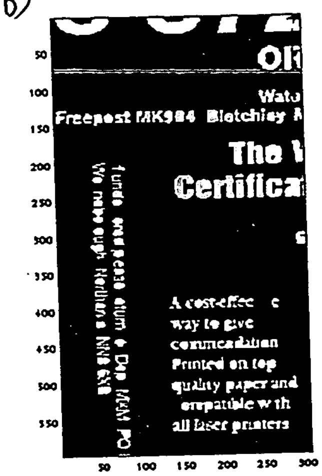
Original
Document Image

a)



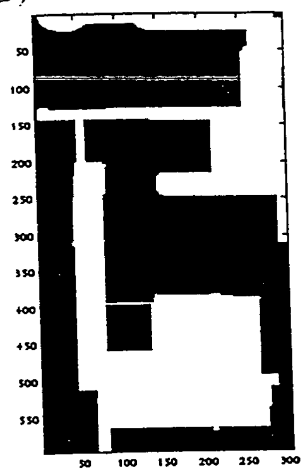
Extracted binary
text objects

b)



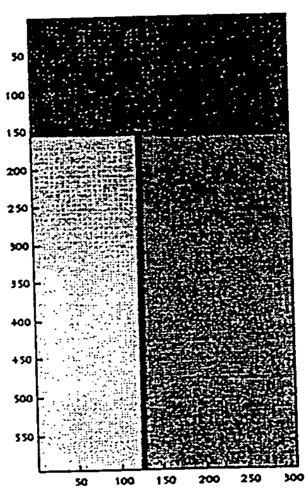
Clustering using simple
merging. All text
objects are clustered
together

c)



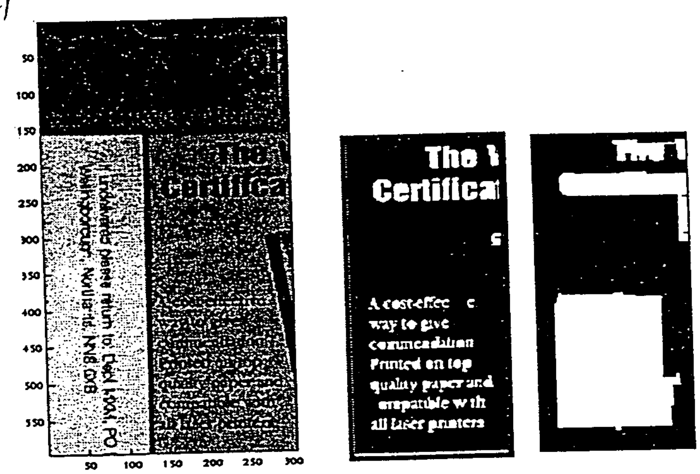
Extract
background from
original

d)



Segment according to background. For each
background region separately apply merging to the
text objects

e)

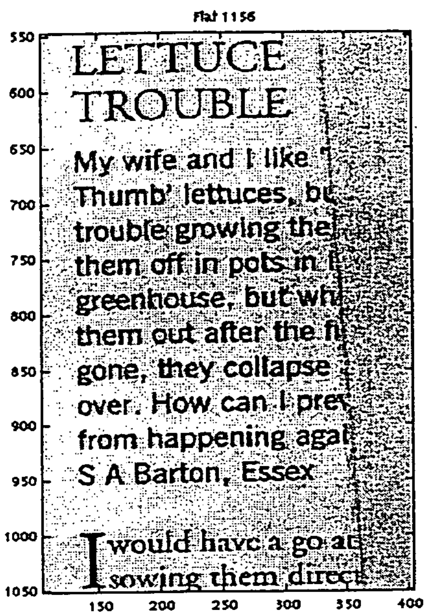


The results are clustered text objects that have
consistent background.

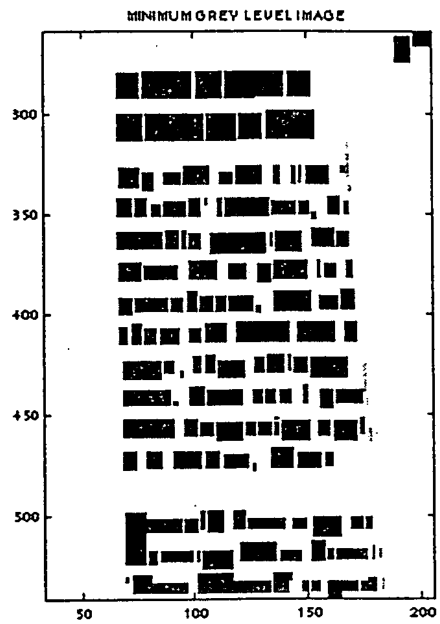
f)

Figure 4.

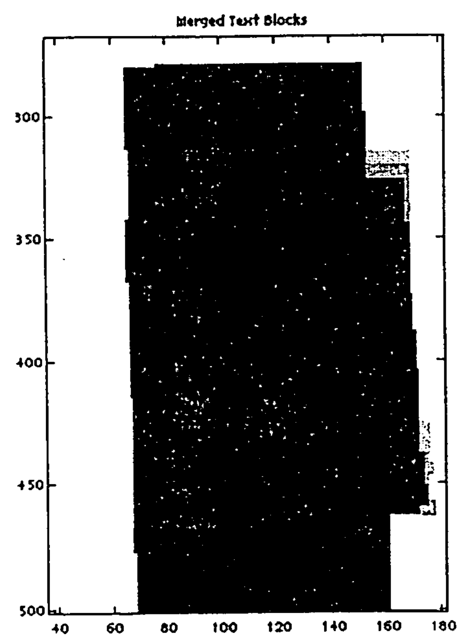
Original Document Image



For each of the binary text objects extracted the local minimum greylevel is obtained from the original document image.
The local minimum greylevel is a measure of the text colour

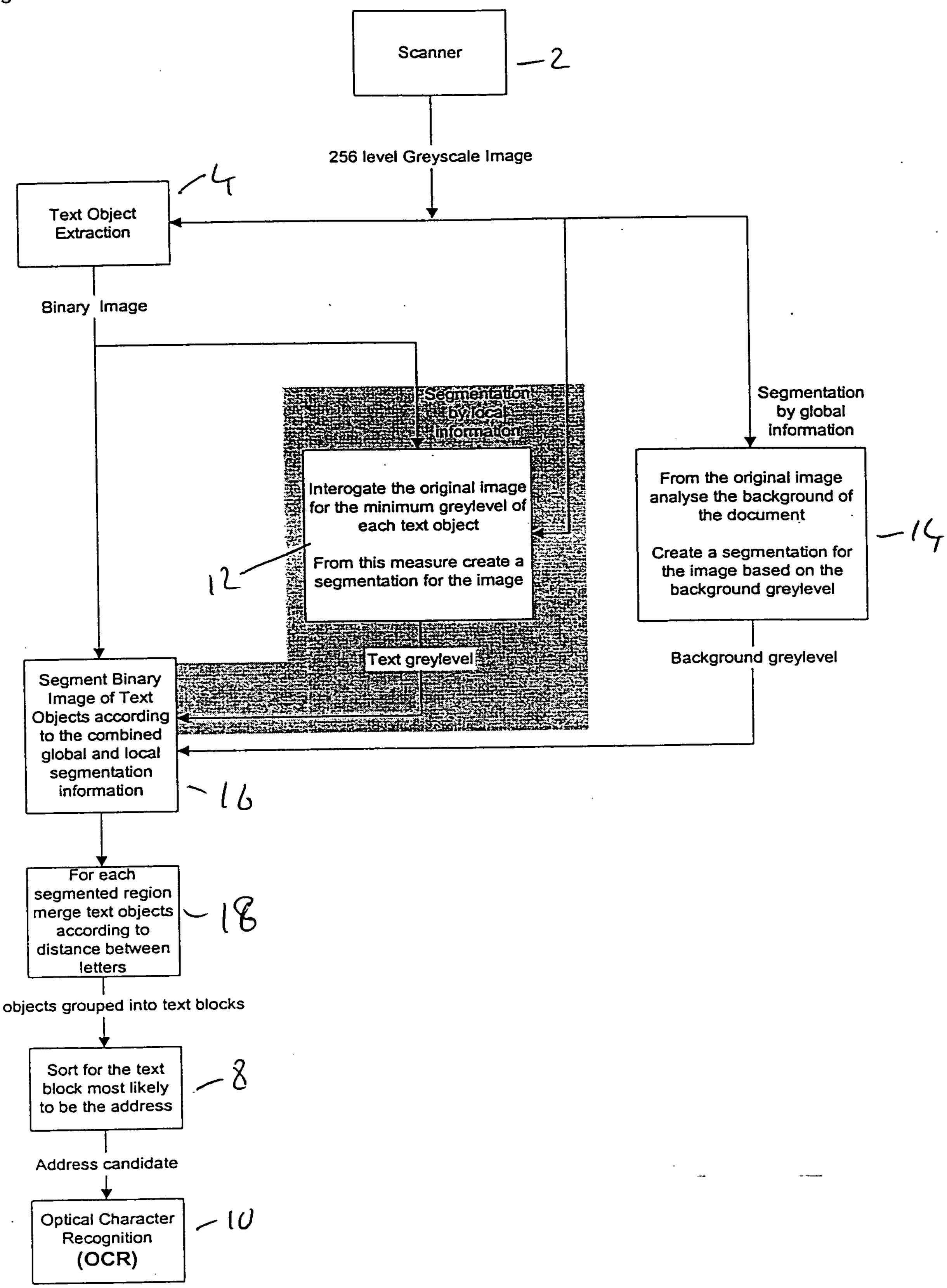


The minimum greylevel data is used to build a segmentation of the image. Each region in segmentation is an area of the image where the text colour is consistent.



5/5

Figure 5.



Scanner
2
256 level Greyscale Image
Text Object Extraction
4
Binary Image
Segmentation by local information
Interogate the original image for the minimum greylevel of each text object
From this measure create a segmentation for the image
12
Segmentation by global information
From the original image analyse the background of the document
Create a segmentation for the image based on the background greylevel
14
Text greylevel
Background greylevel
Segment Binary Image of Text Objects according to the combined global and local segmentation information
16
For each segmented region merge text objects according to distance between letters
18
t objects grouped into text blocks
Sort for the text block most likely to be the address
8
Address candidate
Optical Character Recognition (OCR)
10